

Bayesian Optimization from Human Feedback: Near-Optimal Regret Bounds

ICML 2025



Aya Kayal



Sattar Vakili



Laura Toni




Da-shan Shiu




Alberto Bernacchia

Motivation: Learning from preferences

Which response do you prefer?
Your choice will help make ChatGPT better.

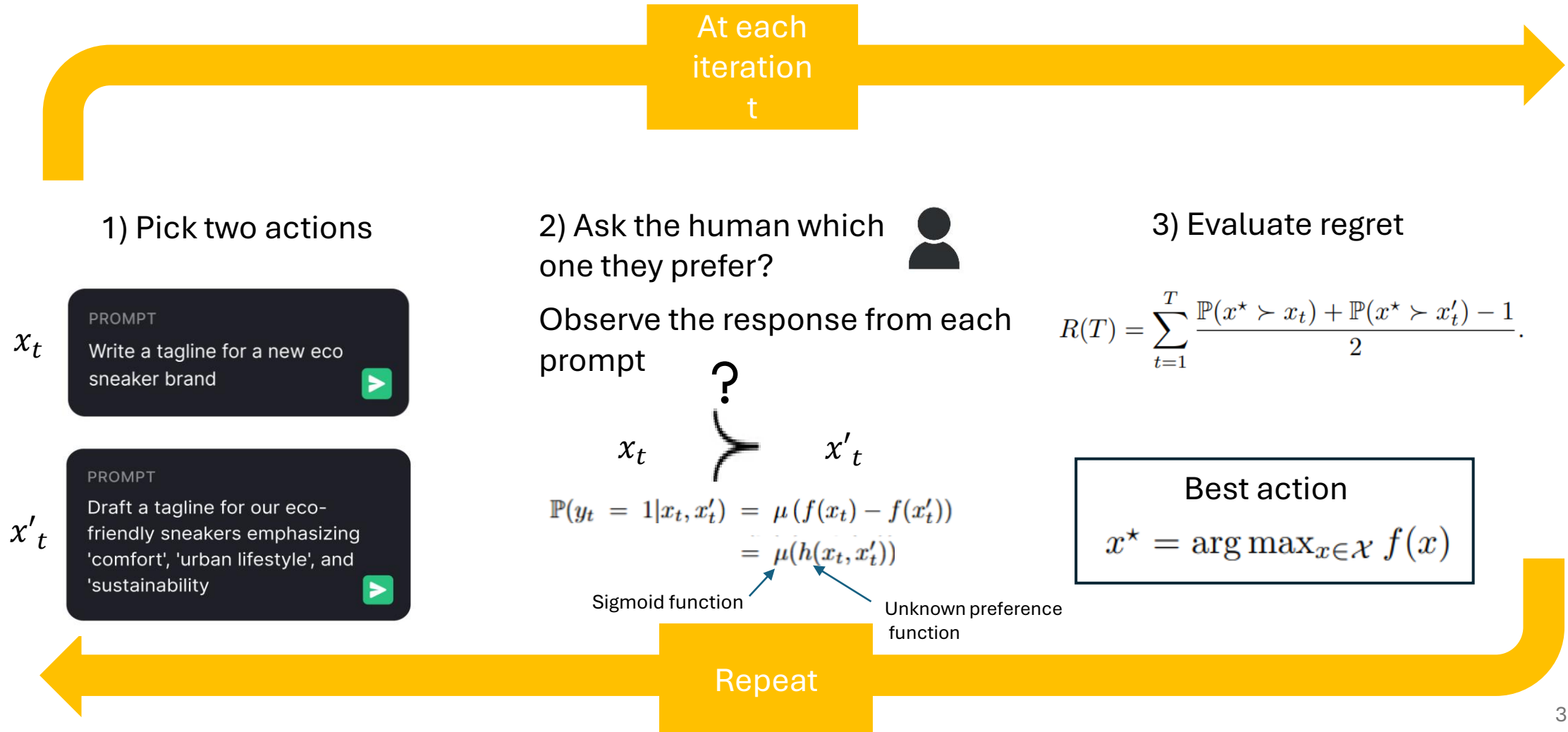
 Response 1
In the OpenAI API, the model you are interacting with is specifically referred to as "gpt-4.0-turbo". This is the latest iteration of the Generative Pre-trained Transformer models developed by OpenAI, offering improved performance and capabilities compared to its predecessors. The "turbo" in the name indicates a more efficient and responsive version of the GPT-4 model, optimized for interactive applications like this one.

 Response 2
In the OpenAI API, the model you are interacting with is referred to as "gpt-4.0-turbo". This identifier is used to specify this particular version and iteration of the Generative Pre-trained Transformer model within the OpenAI API framework.

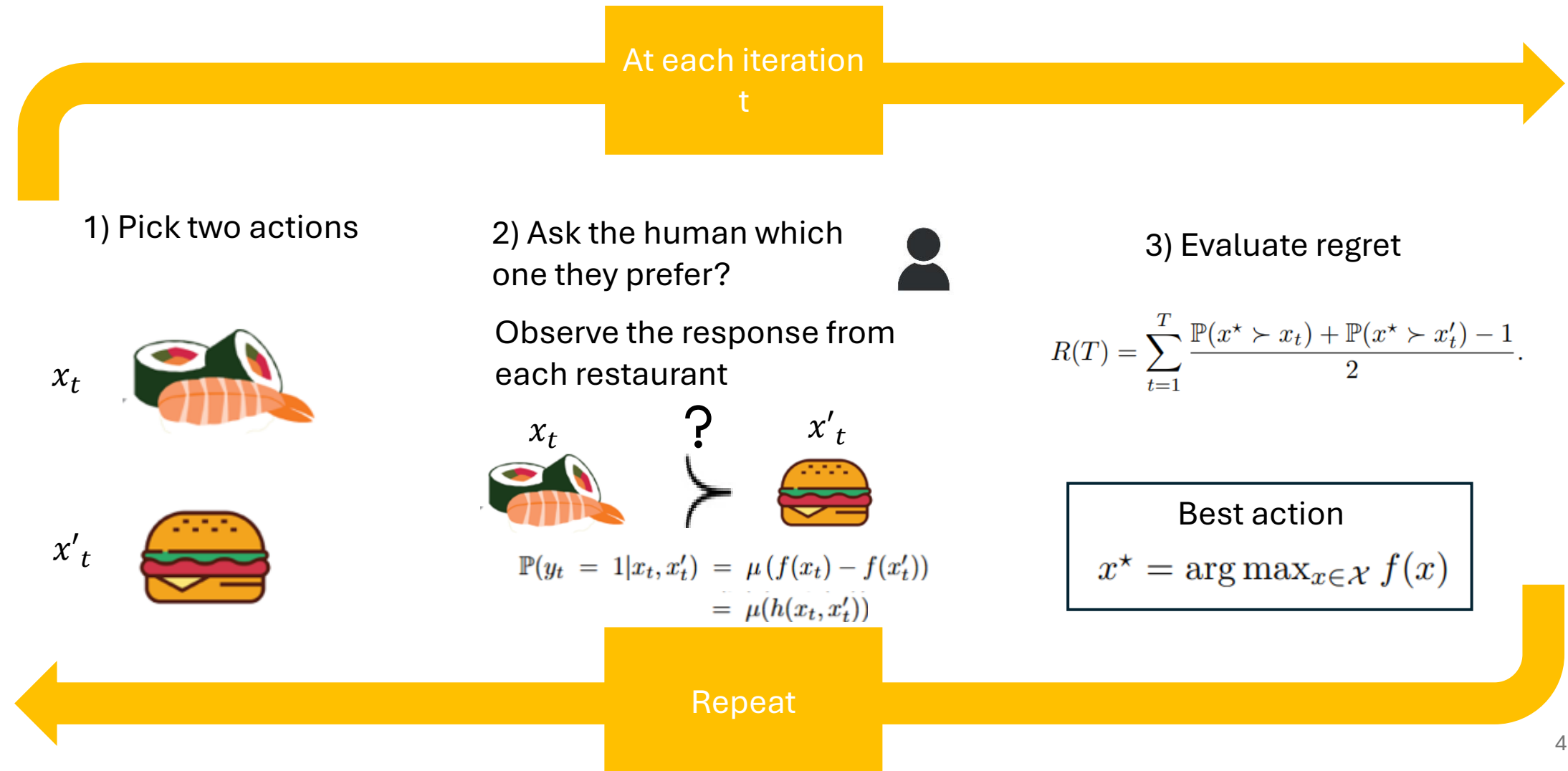
- Humans are often more reliable at providing relative preferences than absolute scores
- Human Feedback is expensive

Goal: How can we learn efficiently using as little preferential feedback as possible?

Prompt optimization can be formulated as Bayesian optimization from Human Feedback (BOHF)



Restaurant recommendation is a BOHF framework



Main Contributions

- **Novel Algorithm:**

We introduce **MR-LPF**, a multi-round structure algorithm for BOHF

- **Improved Regret Bound:**

Prior methods rely on a large constant κ related to the curvature of the link function and a complexity term $\Gamma(T)$ that may grow polynomially with T

Our approach achieves a regret bound of $\tilde{O}\left(\sqrt{\Gamma(T)T}\right)$

- **Empirical Validation:**

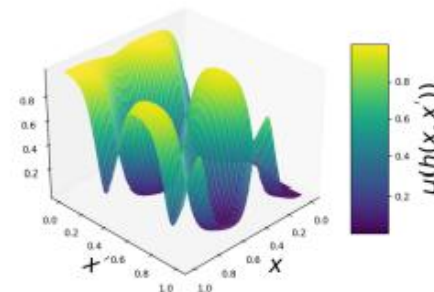
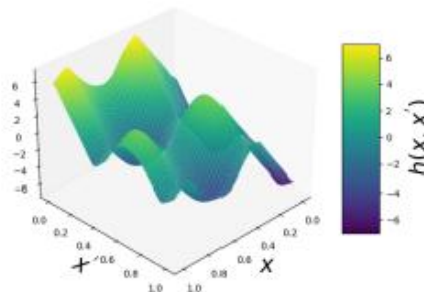
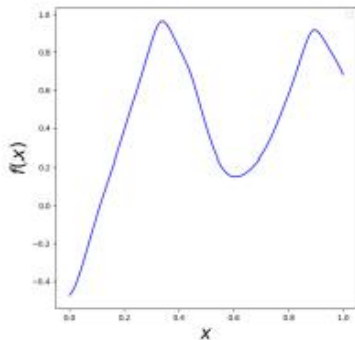
We validate our method through experiments on both **synthetic** and **real-world** datasets.

Preference function estimation $h(x_t, x'_t)$

- The problem resembles a classification-like problem with binary outputs
- Minimize logistic negative log-likelihood loss:

$$\mathcal{L}_{\mathbf{k}}(h, \mathbb{H}_t) = \sum_{i=1}^t -y_i \log \mu(h(x_i, x'_i)) - (1 - y_i) \log(1 - \mu(h(x_i, x'_i))) + \frac{\lambda}{2} \|h\|_{\mathcal{H}_{\mathbf{k}}}^2$$

Apply the representer theorem which provides a parametric representation of h :

$$h_t(\cdot) = \sum_{i=1}^t \theta_i \mathbf{k}(\cdot, (x_i, x'_i))$$


Uncertainty estimation from kernel ridge regression

$$\sigma_t^2(z) = \mathbb{k}(z, z) - \mathbb{k}_t^\top(z)(\mathbb{K}_t + \lambda\kappa I)^{-1}\mathbb{k}_t(z)$$

Where:

- $z = (x, x')$ is a pair of actions
- $\mathbb{k}_t(z) = [\mathbb{k}(z, (x_j, x'_j))]_{j=1}^t$ is the vector of kernel values between the test pair z and observation pairs
- $\mathbb{K}_t = [\mathbb{k}((x_i, x'_i), (x_j, x'_j))]_{i,j=1}^t$ is the kernel matrix over observations pairs
- λ is the regularization coefficient.
- κ depends on the curvature of the link function

Multi-Round Learning from Preference Feedback (MR-LPF)

- Split time horizon T into R rounds with number of samples defined recursively by : $N_0 = \lceil \sqrt{T} \rceil$, $N_r = \lceil \sqrt{N_{r-1} T} \rceil$
- In each round r , for each sample n :
 - Select action pairs with **highest uncertainty** (via kernel ridge regression)
 - Receive **binary feedback**
- Maintain a candidate set of plausible actions
- At the end of each round:
 - **Eliminate actions unlikely to be the best** using upper confidence bound (UCB) of the preference function

MR-LPF acquisition function

$$(x_{(n,r)}, x'_{(n,r)}) = \arg \max_{x, x' \in \mathcal{M}_r} \sigma_{(n-1,r)}(x, x').$$

MR-LPF sub-optimal action elimination

$$\mathcal{M}_{r+1} = \{x \in \mathcal{M}_r \mid \forall x' \in \mathcal{M}_r : \mu(h_{(N_r,r)}(x, x')) + \beta_{(r)} \sigma_{(N_r,r)}(x, x') \geq 0.5\}$$

Analysis

Our regret bounds show an improvement by $\mathcal{O}(\sqrt{\Gamma(T)})$ over prior work, and eliminate the dependence on the curvature of the link function (κ)

Table 1. Comparison of regret bounds in BOHF.

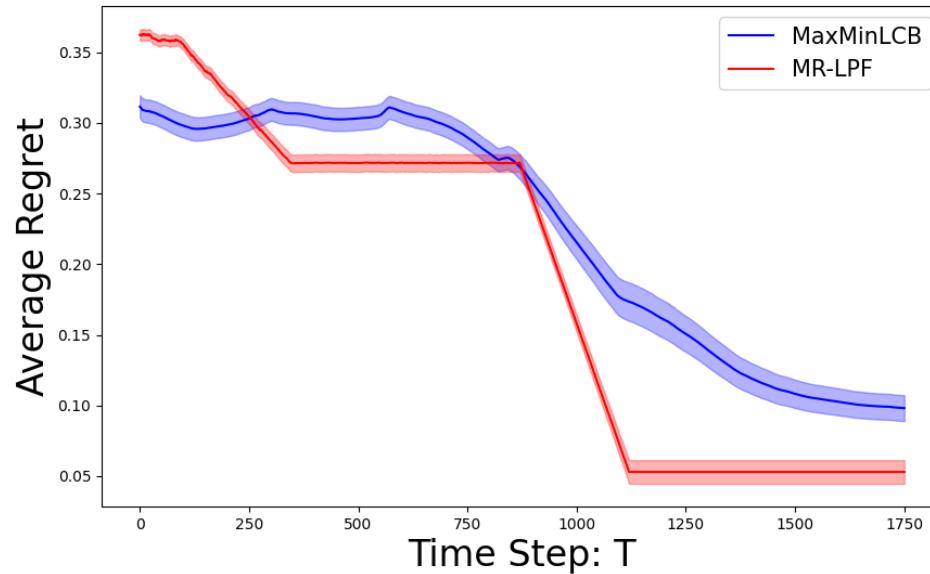
(Pásztor et al., 2024)	(Xu et al., 2024)	This work
$\tilde{\mathcal{O}}\left(\Gamma(T)\kappa^2\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left((\Gamma(T)T)^{3/4}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\Gamma(T)T}\right)$

We match the performance guarantees of conventional BO, despite using a much weaker feedback model.

Pasztor, B., Kassraie, P., and Krause, A. Bandits with preference feedback: A stackelberg game perspective. In Advances in Neural information Processing Systems 38, 2024.

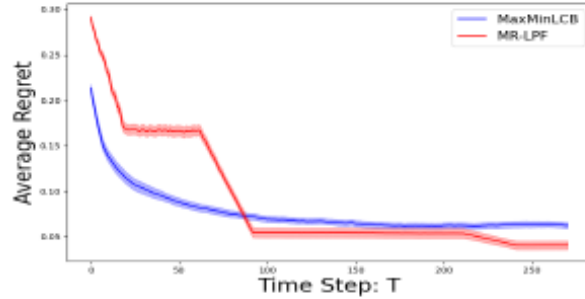
Xu, W., Wang, W., Jiang, Y., Svetozarevic, B., and Jones, C. Principled preferential bayesian optimization. In Forty first International Conference on Machine Learning, 2024.

MR-LPF outperforms the strongest baseline on the YELP restaurant recommendation dataset.

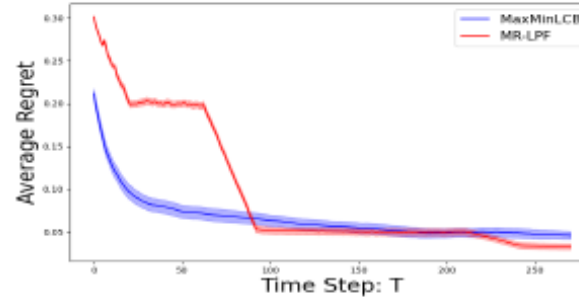


Average regret against T for the experiment with Yelp Open Dataset. The shaded area represents the standard error.

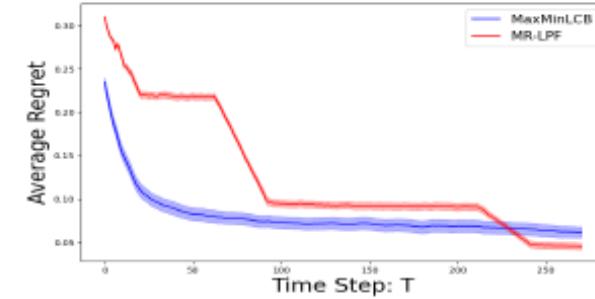
MR-LPF consistently outperforms the baseline across a variety of challenging synthetic test functions



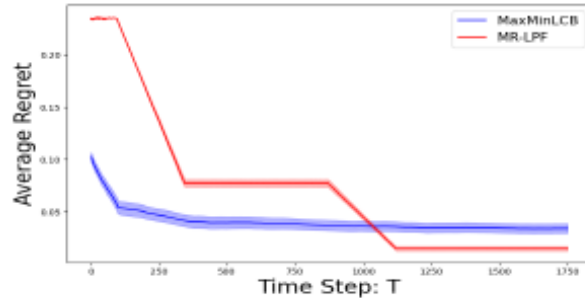
(a) SE kernel (RKHS)



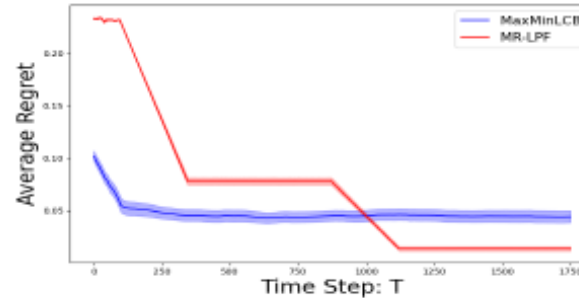
(b) Matérn kernel with $\nu = 2.5$ (RKHS)



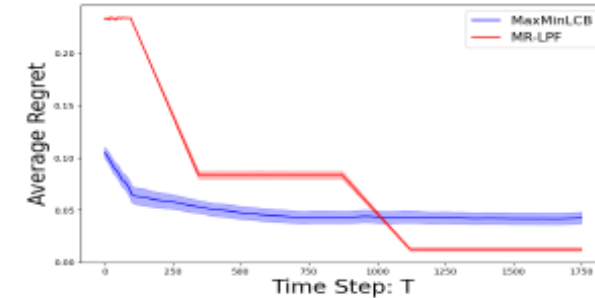
(c) Matérn kernel with $\nu = 1.5$ (RKHS)



(d) SE kernel (Ackley)



(e) Matérn kernel with $\nu = 2.5$ (Ackley)



(f) Matérn kernel with $\nu = 1.5$ (Ackley)

Average Regret against T with RKHS test functions (top row) and Ackley test function (bottom row). The shaded area represents the standard error.

Thank you for your attention !



Scan for the full paper